Data analytics in SDN and NFV: Techniques and Challenges

Girish L
Research Scholar
SIET, Tumkur
Email:girishltumkur@gmail.com

Raviprakash M L
Associate Professor
KIT, Tiptur
Email: raviprakashml@gmail.com

***Abstract*—Software defined networking and network function virtualization are drawing huge attention from researchers both in industry and academia. NFV reduces the capital and operational expenditure of the organization by decoupling the network functions from physical hardware on which they run, which poses new challenges in the perspective of network management such as data management, resource management and performance analysis. Consequentially, novel techniques and strategies are required to address these challenges in efficient way. This paper discusses the most widely used data analytics techniques like machine learning and time series data analysis. Further it describes the review of data mining tools and frameworks. Machine learning helps to overcome the challenges of network management by providing intelligence in network. Hence, in this paper we describe an overview of high level architecture of machine learning analysis framework, the challenges of applying machine learning algorithms in virtual environment and alsosome of the interesting problems of network management which can be solved by using machine learning.**

***Index Terms*—SDN, NFV, Data Mining, Machine Learning.**

I. INTRODUCTION

Computer networks can be divided in three planes of functionality: the data, control, and management planes. Data plane is responsible for forwarding the data, usually handle the end-user generated packets to other end-user devices. The control plane denotes the protocols used to populate the forwarding tables of the data plane devices. The management plane corresponds to the software services, which is used to remotely interact or monitor the network device [1].

In management plane the network policy is defined, the policy enforcement is done by control plane, and the data plane executes it by forwarding data accordingly.

In legacy IP networks, the data and control planes are tightly coupled, embedded in the same networking devices, and the entire structure is highly decentralized. Because of tightly coupled nature, the network is relatively static architecture, difficult to manage and control.

SDN (Software Defined networking) [2] uses the standards-based software abstraction between the control plane and data plane to eliminate the complex and static nature of the legacy distributed network. Software based abstractions solves the architectural problems and makes architecture is more dynamic.

In SDN, the forwarding state in the data plane is managed by a logically centralized control plane. The network devices will acts as simple (packet) forwarding elements by removing the control functionality from them. Also, they are un-awareof the rest of the network and rely on the control plane to update their configuration and forwarding tables.

The network intelligence is moved to a software based external entity: SDN controller, which makes the network is programmable through software program running on top of the controller and it interacts with the data plane devices via a southbound interface (Openflow)[3].

On other hand, with fast growing network usage, the network operators are facing burdens in accommodating the ever-growing demand for new services with diverse proprietary network appliances. Also, it increases both the capital and operational expense of service providers. In typical, NFV is proposed to virtualize the network services that are being carried out by proprietary, dedicated hardware to reduce both CAPEX and OPEX expenditure [4].

The aim of NFV is to address these difficulties by evolving (running) standard network functions like router, firewall, switches and load balancer in a software hosted on virtual machines (VMs) or containers, and that can be moved tovarious locations in the network as required.

The evolution of SDN and NFV networking introduces many new challenges in managing and allocating the network resources in highly dynamically changes system (environment) in efficient manner.

Analytics plays a critical role in managing networks dynamically including self-healing and self-optimization. With predictive analytics can providing actionable intelligence and along with policy decisions, network resources can be dynamically prioritized to meet the changing demands of customers [5].

A good number of research works have come up with many novel algorithms and methods to perform data analytics in SDN and NFV. The main objective is to automate real time decision making at the end which requires machine learning. In this paper, we provide a brief survey of possible different methods of data analysis methods/techniques forthe analytics in software defined networking and network function virtualization.

The section II describes the different data mining techniques for network analytics. Section III describes the main open source tools and framework for data mining. Section IV discusses use the challenges of applying machine learning algorithms in virtual environment and also some of the interesting problems of network management which can be solved by using machine learning.

II. Data Analytics Techniques and Methods

This section discusses about most widely used data anlytics approaches like machine learning and time series data analysis.

A. Machine Learning

Machine learning is an artificial intelligence technique evolved from computational learning theory and pattern recognition [6]. The two important techniques of machine learning are Classification and Clustering. Classification is a function to predict to which class the data object belongs to. Clustering is a function to group a set of objects such that objects belonging to the same group are similar to each other in some manner. Classification is a supervised learning technique. Clustering is an unsupervised learning technique.

Supervised learning algorithms can also be used for regression and anomaly detection. Some of the well-known supervised learning algorithms are Decision tree, Nave Bayes Classification, K-Nearest neighbors, Support Vector Machine, Logistic regression and Linear Regression. Unsupervised algorithms include clustering techniques like K-Means, K-Medoids, DBSCAN and agglomerative.

The task of selecting a specific learning algorithm is acritical step. The features that can be considered in order to se-lect a suitable learning algorithm are accuracy, training speed, linearity and ease of understanding of the algorithm. The table **??** gives the comparison of few widely used supervised and unsupervised algorithms.

Nave Bayes Classification algorithm belongs to a family of probabilistic classifiers based on Bayes theorem with strong independence assumptions between the features. Rich Caruana [8] in his Empirical Comparison of Supervised Learning Algorithms says that Nave Bayes Classification performs better compared to other approaches of classification algorithms like boosted trees or random forests [7].

Decision tree is another classification and regression algorithm which uses tree like model for decision making. It is also referred as CART which means classification and regression tree. The tree includes root node, internal node, leaf node and branches [10]. Root node is a decision making node which splits into two or more mutually exclusive subsets. Each internal node in a tree represents a test on an attribute; each branch represents the outcome of the test on root node and internal nodes. The paths from root node through internal nodes to leaf node represent the classification decision rule. The leaf node represents the final class or label of the decision rule. Decision tree alone does not give good classification performance for complex data but its performance can be improved using hybrid approach i.e. combining decision tree with genetic algorithm [9].

Logistic Regression is a technique for classification not regression. It involves more probabilistic view of classification. It is especially useful in domains where relative probabilities can be miniscule. It gives good accuracy for many simple data sets. It makes no assumptions about distributions of classes in feature space. It can interpret the model coefficients as indicators of feature importance [11].

Linear regression is a linear system in which coefficients can be calculated using linear algebra. It is an algorithm to predict dependent variable based on independent variable [11].

Support Vector Machine algorithm can be used for both classification and regression. It is fast and dependable classification algorithm which performs extremely well for small amount of data. Linear SVM is the newest fast machine learning algorithm used for solving problems on very large datasets. Some of the properties of SVM include flexibility in choosing the similarity function, sparseness of solution when dealing with large data sets, ability to handle large feature spaces and overfitting can be controlled by soft margin approach. SVM is successfully used in many real world applications like, text categorization, image classification, bioinformatics, hand written character recognition etc [12].

K-Nearest neighbors is among simplest of all machine learning algorithms. It is a non-parametric method algorithm used for both classification and regression. In both the cases input consists of k closest points in the feature space. KNN makes predictions using training dataset directly. It is called lazy algorithm because all computations are deferred until classification [13].

B. Time Series Data Analysis

Time series data is a sequence of data objects or events collected at regular time intervals. Data collected at irregular time intervals cannot be considered as time series data. Time series analysis is an application of statistical methods to extract meaningful information and interesting patterns from the data [14].

With time series data we can perform the trend analysis. This helps us to understand and monitor the characteristicsof data that are affecting the trend both on the positive and negative aspects, which allows us to change the outcome by carefully controlling parameters of data objects. Future trend prediction is gaining attention in many industry applications and also it is highly active filed in the academia too.

For the building time-series modelling most of the applications uses linear models. There are three kinds of linear models: moving-average (MA), autoregressive (AR) and ARMA models.

AR models are widely used models for time series, within the framework of least-squares regression they can be fully estimated and tested.

For modeling univariate time series, the moving-average (MA) model is used in most of the applications. This model

Comparison of supervised and Unsupervised learning algorithms					
Algorithm	Accuracy	Training Speed	Linear/Non Linear	Ease of Understanding	Notes
Decision tree	Medium	Fast	Non Linear	Moderate	For prediction it is not used on its own because its too simple and is not good enough for complicated data.
Nave Bayes Classification	Medium	Fast (Excluding feature extraction)	Linear	Moderate	It is based on Bayes' theorem. Assumes independence among predictors. Use full for large datasets.
K-Nearest neighbors	Medium	Moderate	Linear	Easy	Non parametric lazy learning algorithm
Support Vector Machine	High	Moderate	Linear	Moderate	Good for large feature sets.
Logistic regression	Medium	Fast	Linear	Moderate	It is simple for capturing complex relationships between variables.
Linear Regression	Medium	Fas	Linear	Easy	Sometimes too simple to capture complex relationships between variables.
K-means	Medium	Moderate	Non Linear	Moderate	A hard-margin, geometric clustering algorithm, where each data point is assigned to its closest centroid.

TABLE I
COMPARISON OF SUPERVISED LEARNING ALGORITHMS

states that the output variable depends linearly on the historic and present values of the stochastic term.

Another method for modeling of time series forecasting [15] is autoregressive integrated moving average (ARIMA) model. This model is a generalization of an autoregressive moving average model which in turn is a combination of the moving average and the auto-regressive models.

To obtain most accurate forecasting model, few researchers have been used artificial neural network as a substitute to these linear model-driven approaches. Due to their features artificial neural networks are belonging to the data-driven approaches [16]. So, the data analysis is depends on the available data and also with relationships among the variables.

Data in time series has a lot of variations, therefore, both clustering and classification methods are used. Classification methods includes decision tree and SVM (support vector machine).

Clustering high dimensionality and large scale time series datasets become a great challenge because it is required an interactive analysis in real practice. Three clustering algorithms are used for large time series data: CLARANS, DBSCAN and DENCLUE. CLARANS is an improved k-medoids method that is more effective and efficient. DBSCAN, a Density-based algorithm treats clusters as dense regions of objectsin the spatial space separated by regions of low density [17].DENCLUE is a highly efficient density-based algorithm to deal with high dimensional datasets and it is faster than CLARANS and DBSCAN.

III. TOOLS AND FRAMEWORKS

This section decsribes the various open source tools of data mining with an comparision review. Also, discusses about the different machine learning frameworks.

A. Open Source Tools for Data Mining

There are many data mining tools for handling large datasets. The main open source tools are scikit-learn, R, WEKA, KNIME, Orange, MOA, RapidMiner and SAMOAas shown below.

Scikit-Learn is an open source machine learning tool written in python. It includes algorithms for classification, regression and clustering like SVM, gradient boosting, random forests, k-means and DBSCAN. It is designed to work with python numerical and scientific libraries NumPy and SciPy [18],[20].

Weka is an open source tool for machine learning written in Java, developed at University of Waikato, New Zealand. Because it is written in Java, it runs on any modern computing platform. Also it supports data mining methods like classification, clustering, regression, feature selection and visualization. It is more suitable for classification and regression problems and less with descriptive statistics and clustering methods [19],[20].

Rapid-Miner is a general data mining tool developed by Rapid I, a company founded by Ralf Klinkenberg and Ingo Mierswa. It provides user friendly GUI. It supports various machine learning tasks like data loading and transformation, ETL, (Extract, transform and load), data pre-processing, predictive analytics, statistical modelling, evaluation and visualization [20],[21],[22].

R is an interpreted open source programming language and software environment for statistical computing and graphics. It runs on variety of LINUX platforms, Windows and Mac OS. R supports a series of graphical and statistical techniques like linear and nonlinear modelling, classification, clustering, and Time series analysis. It supports algorithms like Decision tree, k-means, and x-means and fuzzy c-means clustering, Nave Bayes function based learning etc.[21].

KNIME (Konstanz Information Miner) is an open source, reporting and integration platform. It is written in Java and based on eclipse. KNIME version 2.1 is released under GPLv3 with an exception that allows others to use the well-defined node API to add proprietary extensions. Through its modular data pipelining concept it can be combines with different components for data mining and machine learning techniques [23].

	Language	Main purpose	User Interface	Algorithms
Scikit-Learn	Python	Machine learning package.	Command Line	logistic regression and linear, SVM
Weka	Java	Data Mining	Graphical User Interface and Command Line	Decision tree, DBSCAN, k-means
Rapid Miner	Java	General Data Mining	Graphical User Interface	CluStream for clustering, Decision rule learner for regression, Vertical Hoeffding Tree (VHT) an extended version of streaming decision tree for classification.
R	R, C++ , Fortran	Computation and Statistics	Graphical User Interface and Command Line	Decision tree learner classification rules Hierarchical clustering, Density Based Clustering DBSCAN, Time Series analysis Data Visualization.
KNIME	Java	General Data Mining	Graphical User Interface	Decision tree learner classification rules, Hierarchical clustering, Density Based Clustering DBSCAN, Partition based clustering k-means, Time Series analysis Data Visualization.

TABLE II
OVERVIEW OF OPEN SOURCE TOOLS

B. Machine learning frameworks

Although there are many machine learning frameworks available, there are few top contenders which stands out, four of which are: Google Tensorflow, Microsoft CNTK, Apache MXNet, and Berkeley AI Research Caffe.

1) Tensorflow: Tensorflow, which is an open-source machine learning framework developed by Google, with a goal of creating a uniform way of producing deep learning re-search or products. It works by utilizing symbolic creationof computation graphs and has both a Python, C++, and aJava implementation. It is worth noting that despite a small core written in C++, much of the functionality is written in Python. Another huge benefit of Tensorflow is Tensorboard, which is a powerful tool that allows the user to visualize manydifferent aspects of your model and your data, allowing you tocreate the proper type of network and optimize it much more quickly and effectively than many frameworks [24],[25].

2) CNTK: CNTK, is an open-source machine learning network developed by Microsoft, which has quite a long history, originally started by researchers working on voice recognition technologies, although it has since expanded to a general use framework. It works similarly as tensorflow, utilizing symbolic the creation of computation graphs, and is arguably even easier to run on a distributed network in the cloud, simply requiring a few lines of code. It also runs a significant amount faster and with a higher degree of accuracy than Tensorflow, since the core and the functionality of the framework is all written in C++, which is more efficient at data processing. It is important to note that there is also a python framework for CNTK as well [24].

3) MxNet: MXNet, which is an open-source framework developed by a variety of industry and university partners, and is currently in the apache incubator program. It offers a unique set of advantages over both Tensorflow and CNTK. First of all, it has native support for a wider range of languages Python, C++, R, and Scala. It also allows for different parts of your models to use either symbolic or imperative programming, which neither Tensorflow nor CNTK allows. MXNet has poor support for recurrent neural networks (RNNs), which are great for things like pattern recognition and prediction of temporal data. It also has a smaller community for support, an area which Tensorflow beats both MXNet and CNTK [26].

4) Caffe: The last framework to be discussed is Caffe, an open-source framework developed by Berkeley Artificial Intelligence Research (BAIR). It was primarily built for computer vision applications, which is an area which still shines today. It is also fairly fast, being written in C++. One notable feature of Caffe is that it was designed with ease-of-use in mind, making many key functions implementation not take more than a few lines of code, such as switching from training on a GPU to a CPU and vice versa [27].

Overall, all these frameworks Tensorflow, CNTK, MXNet, and Caffe: all deserve their place on this list and each offer their own unique pros and cons. Right now, it seems that Tensorflow has the largest user pool and is being developed more quickly than some of the others, and therefore also has a lot of community support along with state-of-the-art features, making it powerful for both research purposes and industry.

IV. MACHINE LEARNING IN SDN AND NFV

This section describes the challenges of applying machine learning algorithms in virtual environment and also some of the interesting problems of network management which canbe solved by using machine learning.

A. Architecture

High level architecture of the automated machine learning analysis framework with an overview is presented in the Figure 1. The proposed architecture is enhanced to both batch and real time machine learning solutions to enable bigdata analytics [28]. The data is collected from NFV environment and systems that consume resources offered by NFV framework. Collected data will be stored in storage devices, later can develop various analytical applications like dynamic resource allocation, failure prediction and performance degradation [29],[30].
The proposed architecture consists of following sub components:

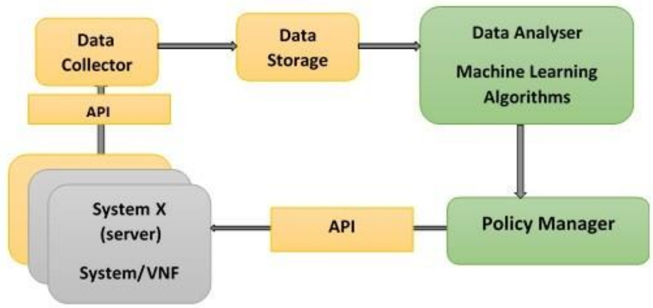

Fig. 1. Architectural Overview.

Data Collection: Monitoring softwares like monasca collecting the data from multiple sources from an NFV frameworks and also with systems that consumes resources offered by framework.

Data storage: It is used for storing real time and historical data and makes them accessible for different components of data analyser. Example for data storage is influxdb.

Machine learning algorithms can be used for different needs, such as, dynamic resource allocation, performance degradation analysis, failure prediction and then adjust resource offering accordingly. The given output is pass to policy managers for policy generation.

Policy managers: This component is mainly responsible for translating obtained outputs to policies that can be directly understandable and also it gives the recommendation regarding policies to the NFV architectural framework.

B. Challenges

Meeting physical and virtual network requirements:

SDN and NFV are being deployed in existing networks which means analytics solutions must be aware of both virtual and legacy environment. Data collection in such environment is challenging one and also should consider both physical and virtual network requirements.

Complex Modeling: It is complex to build mathematical modeling for dynamically changing distributed network system and for creating the actionable analytics for such network configurations is also challenging.

Monetizing the data: As the data is collectd from multiple sources is virtual environment, it is important to extract the data into meaningful formats and to monetize the data before applying algorithms.

Data management: Huge amount of data will be generated within NFV and SDN environment, which will need to be efficiently stored, indexed and leveraged for achieving insights into network and user behavior.

C. Applications

By using machine learning in virtual environment, we can solve some of the interesting problems like:

Event Correlation analysis. Event correlation techniques have been used as an important method in networks to detect reasons of performance degradation problems. Correlation can be applied different types of the data in virtual environment. In virtual environment correlation analysis can be apply for infrastructure metrics/events and test performance metrics to understand the performance behaviour.

Trend Analysis The trend analysis can be used to identify the trends in the performance of both physical and virtual network elements. The identified trend can be used to predict the future network statistics by using supporting machine learning algorithms over a period of time.

Anomaly detection. Using machine learning algorithms can detect the abnormal behaviour of services and hosts by analysing metrics and events collected from all the layers of cloud infrastructure.

Failure prediction: A failure prediction system can be used in networking to avoid the unexpected failure in advance.Based on the machine learning techniques, the failure predictor analysis can find the failure before it happen and send the appropriate alert to the network operator.

V. CONCLUSION

The SDN and NFV networking will pose a wide set of new challenges for network management and analytics. This paper describes and reviews machine learning, its challenges, different techniques and open source tools for analysis of network data. In addition to that paper discusses on how to apply machine learning techniques in NFV and SDN networking. Hence, in this paper we describe an overview of high level architecture of machine learning analysis framework, the challenges of applying machine learning algorithms in virtual environment and also some of the interesting problems of network management which can be solved by using machine learning.

REFERENCES

[1] Langston Nashold,Rayan Krishnan, ”using LSTM and SARIMA Models to Forecast Cluster CPU Usage”.published:16 Jul 2020.

[2] Pranav T P, Charan S, Darshan M R, “Devops Methods for Automation of Server Management using Ansible ”, International Journal of Advanced Scientific Innovation, vol. 1, no. 2, pp. 7-13, May 2021.

[3] Sak, Has¸im, Andrew Senior, and Franc¸oise Beaufays. ”Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition.” arXiv preprint arXiv:1402.1128 (2014).

[4] Jitendra Kumar,Rimsha Goomer,Ashutosh Kumar Singh,”Long Short Term Memory Recurrent Neural Network(LSTM-RNN)Based Workload Forecasting Model For Cloud Datacenters”.Published:7-8 Dec 2017.

[5] Z. Asha Farhath,B.Arputhamary,Dr.L.Arockiam,”A Servey On ARIMA Forecasting Using Time Series Model”.Published:8 Aug 2016.

[6] KUMAR Manoj,ANAND Madhu,”An Application Of Time Series ARIMA Forecasting Model For Predicting Sugarcane Production In India”.

[7] Ratnadip Adhikari,R.K. Agarwal,”An Introductory Study On Time Series Modeling And Forecasting”.DOI:10.13140/2.1.2771.8084.

[8] Bijeshdhyani,Manish Kumar,Poonam Verma,Abhishik Jain,”Stock Market Forecasting Technique Using ARIMA Model”. Mar 2020.

[9] L. Girish and S. K. N. Rao, "Mathematical tools and methods for analysis of SDN: A comprehensive survey," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 2016, pp. 718-724, doi: 10.1109/IC3I.2016.7918055.

[10] Anupama K C,Shivakumar B R,Nagaraja R,"Resource Utilization Prediction in Cloud Computing using Hybrid Model".Vol 12.No.4,2021.

[11] Khanum, Salma, and L. Girish. "Meta heuristic approach for task scheduling in cloud datacenter for optimum performance." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 4.5 (2015): 2070-2074.

[12] [12] Girish, L., Rao, S.K.N. Anomaly detection in cloud environment using artificial intelligence techniques. Computing (2021).https://doi.org/10.1007/s00607-021-00941-x

[13] Prajwal S, Siddhartha M, Charan S. (2021). DDos Detection and Mitigation SDN using support vector machine. International Journal of Advanced Scientific Innovation, 1(2), 26-31. https://doi.org/10.5281/zenodo.4782280

[14] Girish L, Rao SKN (2020) Quantifying sensitivity and performance degradation of virtual machines using machine learning. J Comput Theor Nanosci. https://doi.org/10.1166/jctn.2020.9019

[15] Rashmi TV, Prasanna MK, Girish L (2015) Load balancing as a service in Openstack-Liberty. Int J Sci Technol Res 4(8):70–73

[16] Raviprakash M L, C S Pillai, Ananda babu J, "Object Recognition using Lucas-Kanade Technique and Support Vector Machine Based Classification in Video Surveillance Systems", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

[17] Ravi Prakash M L , Chethan Chandra S Basavaraddi , Ananda Babu J, "Face Recognition Using Hybrid Neuro Fuzzy Network for Occluded Images", International Journal of Science and Research (IJSR), 2018

[18] MANOJ KUMAR D P. (2021). A Clustering and Detection of Liver Disease in Indian Patient Using Machine Learning Algorithms. SPAST Abstracts, 1(01). Retrieved from https://spast.org/techrep/article/view/838

[19] Chethan Chandra S Basavaraddi, Raviprakash M L, Dr. C S Pillai, "Face Recognition from Feed Forward Neural Network Using Occluded Images for Automating the Surveillance Using Hybrid Neuro Fuzzy Network", International Journal of Engineering Applied Sciences and Technology, 10.33564/ijeast.2020.v05i02.087, 2020, Vol 5 (2) pp. 508-519

[20] Diego Kreutz, Paulo Esteves Verssimo and Fernando M. V. Ramos, Software-Defined Networking:A Comprehensive Survey Proceedings of the IEEE, Vol. 103, No. 1, January 2015.

[21] Yosr Jarraya,Taous Madi and Mourad Debbabi, A Survey and a Layered Taxonomy of Software-Defined Networking IEEE COMMUNICATION SURVEYS and TUTORIALS, VOL. 16, NO. 4, FOURTH QUARTER 2014.

[22] Nayana Y, Mr.JustinGopinath, Girish.L *DDoS Mitigation using Software Defined Network* International Journal of Engineering Trends and Tech- nology (IJETT) Volume 24 Number 5- June 2015.

[23] Albert Mestres, Alberto Rodriguez-Natal, Josep Carner, Pere Barlet-Ros, Eduard Alarc on, *Knowledge-Defined Networking* ACM SIGCOMMComputer Communication Review, Volume 47, Issue 3, July 2017.

[24] Hoda A. Abdel Hafez, *Mining Big Data in Telecommunications Industry: Challenges, Techniques, and Revenue Opportunity* World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:10, No:1, 2016.

[25] M. Martinez-Arroyo and L. E. Sucar, *Learning an Optimal Naive Bayes Classifier* 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 2006, pp. 958-958.doi: 10.1109/ICPR.2006.749.

[26] Rich Caruana and Alexandru Niculescu-Mizil, *An empirical comparison of supervised learning algorithms* In Proceedings of the 23rd international conference on Machine learning (ICML 2006). ACM, New York, NY, USA, 161-168. DOI: https://doi.org/10.1145/1143844.1143865.

[27] Deborah R. Carvalho, Alex A. Freitas, *A hybrid decision tree/genetic algorithm method for data mining* Information Sciences,Volume 163, Issues 13,2004.

[28] Yan-yan SONG, Ying LU, *Decision tree methods: applications for classification and prediction* Shanghai Arch Psychiatry. 2015 Apr 25;27(2):130-5. doi: 10.11919/j.issn.1002-0829.215044.

[29] B. Pavlyshenko, *Machine learning, linear and Bayesian models forlogistic regression in failure detection problems* 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 2046-2050.